FETNet: Feature Exchange Transformer Network for RGB-D Object Detection

Zhibin Xiao[1, 3]
xzb18@mails.tsinghua.edu.cn
Jing-Hao Xue[2]
jinghao.xue@ucl.ac.uk
Pengwei Xie[1]
xpw18@mails.tsinghua.edu.cn
Guijin Wang[1]
wangguijin@tsinghua.edu.cn

[1] Department of Electric Engineering
Tsinghua University
Beijing, China

[2] Department of Statistical Science
University College London
London, U.K

[3] Tsinghua Shenzhen International
Graduate School
Tsinghua University
Shenzhen, China

Abstract

In RGB-D object detection, due to the inherent difference between the RGB and Depth modalities, it remains challenging to simultaneously leverage sensed photometric and depth information. In this paper, to address this issue, we propose a Feature Exchange Transformer Network (FETNet), which consists of two well-designed components: the Feature Exchange Module (FEM), and the Multi-modal Vision Transformer (MViT). Specially, we propose the FEM to exchange part of the channels between RGB and depth features at each backbone stage, which facilitates the information flow, and bridges the gap, between the two modalities. Inspired by the success of Vision Transformer (ViT), we develop the variant MViT to effectively fuse multi-modal features and exploit the attention between the RGB and depth features. Different from previous methods developing from specified RGB detection algorithm, our proposal is generic. Extensive experiments prove that, when the proposed modules are integrated into mainstream RGB object detection methods, their RGB-D counterparts can obtain significant performance gains. Moreover, our FETNet surpasses state-of-the-art RGB-D detectors by *7.0% mAP* on SUN RGB-D and *1.7% mAP* on NYU Depth v2, which also well demonstrates the effectiveness of the proposed method.

1 Introduction

Object detection, which aims to locate and classify objects from input images, is a fundamental yet challenging task in computer vision. Remarkable progress has been made in this field, benefiting many intelligent tasks, including autonomous driving, vision navigation, and scene understanding. With the rapid development of commercial depth sensors, depth images can be readily collected, which are also expected to complement RGB images to promote object detection performance.

Consequently, RGB-D object detection has attracted increasing attention in the past few years. Early works [1, 2, 3] take depth maps as the fourth channel of the corresponding RGB images, and use various hand-designed kernel descriptors to model object size and 3D shape. However, since RGB information and depth information are inherently different, it is difficult to effectively extract features from the simply concatenated data. To alleviate this problem, recent works [10, 15, 16, 18, 31] have adopted two parallel backbone networks to extract RGB-D features separately. RGBD R-CNN [15] generalizes the R-CNN detector [13] into a two-stream network for the RGB and depth modalities. Large-scale CNNs pre-trained on RGB images are used to help extract depth features. To initialize the depth network with better parameters, Gupta *et al.* [16] transfer supervision from the large-scale labeled RGB modality to the unlabeled paired depth modality. Xu *et al.* [31] propose a correlated detection module to mitigate the disagreements between the modality-specific results. However, these methods have insufficient ability to learn long-range attention between the two modalities, which limits their performance. Li *et al.* [18] propose a cross-modal attentional context framework to incorporate the correlated information from different modalities. But the feature information flow between the two modalities at the backbone stage has not received sufficient attention in these methods, which hinders the backbone network from extracting effective modality-specific features. In addition, most existing methods are specially designed for R-CNN [13] and Fast R-CNN [12], which prevents them from exploiting the further development in the RGB-based object detection field.

Aiming to address the above-mentioned issues in RGB-D object detection, we propose a novel Feature Exchange Transformer Network (FETNet), which is independent of specific RGB detectors. It consists of two well-designed components. First, we introduce the Feature Exchange Module (FEM) to partially swap the RGB and depth features at each backbone stage, which facilitates the information flow between the two modalities. Second, inspired by the recent success of Transformer [28], we develop the Multi-modal Vision Transformer (MViT) to learn the global and local attention of the two modalities and perform multi-modal feature fusion. Embedded with these two new modules, FETNet surpasses state-of-the-art RGB-D detectors by **7.0%** *mAP* on SUN RGB-D and **1.7%** *mAP* on NYU Depth v2. Furthermore, by integrating the proposed modules, various RGB object detectors can be extended to their RGB-D variants, and significant performance gains can be obtained. These clearly demonstrate the value and versatility of the proposed modules.

2 Related Work

2.1 Depth Feature Extraction

Depth maps, which contain information related to the distance from specified viewpoints to the surfaces of scene objects, are inherently different from RGB images of the scene. Therefore, it is challenging to extract features from the distance-related modality effectively.

Early works [1, 2, 3] mainly focus on the hand-designed operators. By taking depth maps as an extra channel of corresponding RGB images, Bo *et al.* [3] leverages hand-designed features such as SIFT and multiple shape features from the depth channel. To facilitate the depth feature extraction, Gupta *et al.* [15] propose a geocentric embedding to convert single-channel depth maps into three-channel HHA format (Horizontal disparity, Height above ground, and Angle with respect to gravity direction). The HHA format is adopted by some following works [18, 31]. However, it introduces a hand-designed conversion, and the conversion process is time-consuming [17]. In this paper, we shall demonstrate that the handcrafted conversion is unnecessary, and we can achieve better performance when taking raw depth maps as input.

Recent works [10, 15, 16, 18, 31] have been dedicated to extracting depth features utilizing ImageNet-pretrained CNNs. To bridge the gap between the RGB-pretrained CNN and input depth data, Gupta *et al.* [16] train the depth backbone by teaching the network to reproduce the mid-level semantic representations learned from well-labeled RGB counterparts. But the information flow between depth and RGB features is blocked in these methods, which hinders the backbone network from learning modality-specific representations.

2.2 RGB-D Information Fusion

Many algorithms dedicate to fusing RGB-D features have been proposed. RGBD R-CNN [15], a two-stream network extended from R-CNN [13], uses two large-scale CNNs pre-trained on RGB images to extract RGB-D features separately, and fuses multi-modal features at a late stage. Li *et al.* [18] develop a cross-modal attentional context network by generalizing Fast R-CNN [12], and introduce LSTM [14] to recurrently generate contextual information from both RGB and depth data. On the basis of Faster R-CNN [24], Xu *et al.* [31] propose a modality-correlated and modality-specific detection network. They introduce a third subnet to learn modality-correlated representations from the modality-specific RGB and depth backbone features at early stages, to mitigate the disagreements between the results from different modalities. However, these methods rely on specific RGB detection algorithms, which limits their versatility.

Different from the above-mentioned algorithms, the proposed modules (*i.e.*, FEM and MViT) are integrated into the backbone network and the widely used feature pyramid layer, respectively. This ensures the proposed method to be independent of any specific RGB object detection framework, and boosts the performance of various object detection methods.

3 Method

3.1 Network Overview

Fig. 1 illustrates the architecture of the proposed FETNet. It takes the RGB images and the corresponding depth maps as input. The RGB images and the depth maps are fed into two different backbone networks. At each backbone stage, we integrate a Feature Exchange Module to partially swap the RGB-D features. MViT takes the exchanged features at each same stage as input, to capture their global and local attention and conduct multi-modal feature fusion. Output features of each level in MViT are fed into a weight-shared head to locate and classify objects in the images. Finally, after Non-Maximum Suppression (NMS), we obtain the final detection results.

3.2 Feature Exchange Module (FEM)

Due to the gap between photometric (RGB) and geometric information, it is tricky to fuse RGB-D features directly. Motivated by the Temporal Shift Module [21] that moves feature maps along the temporal dimension, we introduce the Feature Exchange Module (FEM) that partially exchanges the RGB features with the depth features at each backbone stage to add

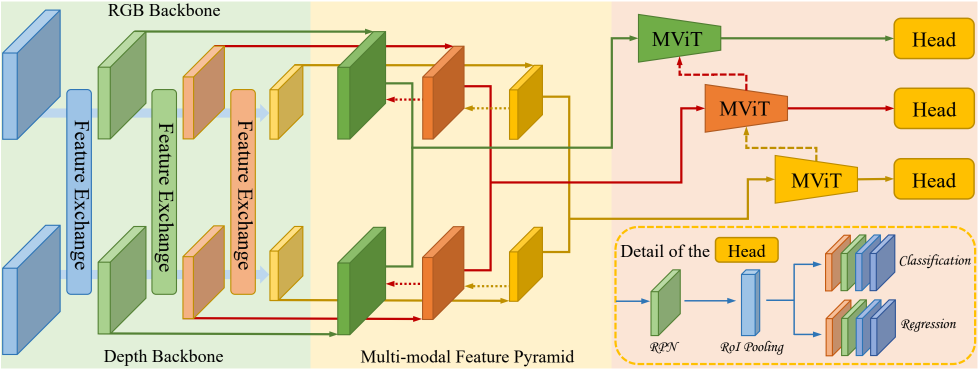

Figure 1: Network architecture of the proposed FETNet. The colored frames between the backbone blocks are the Feature Exchange Modules. Features from each *MViT* are fed into a weight-shared *Head* to obtain the location and classification results. For simplicity, we only show the details of one *Head* in the dashed box.

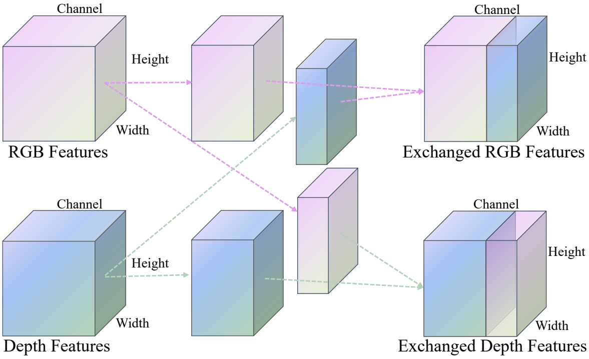

Figure 2: Schematic diagram of the proposed Feature Exchange Module (FEM). We partially exchange RGB-D features with Depth features at each backbone stage to build an information flow between them.

an information flow between them. It thus bridges the gap between these two modalities. As shown in Fig. 2, we split RGB features $\mathcal{F}_{rgb}$ and Depth features $\mathcal{F}_{depth}$ of C channels into two blocks ($\mathcal{S}_k$ with $k \times C$ channels and $\mathcal{S}_{1-k}$ with $(1-k) \times C$ channels, with $k \in [0,1]$), respectively. Then we exchange the block in the RGB features with the corresponding block in the Depth features.

The proposed FEM can be formulated as

$$\begin{aligned} \mathcal{F}^{i+1}_{rgb} &= \mathcal{C}at(\mathcal{S}_k(\mathcal{F}^i_{depth}), \mathcal{S}_{1-k}(\mathcal{F}^i_{rgb})), \\ \mathcal{F}^{i+1}_{depth} &= \mathcal{C}at(\mathcal{S}_k(\mathcal{F}^i_{rgb}), \mathcal{S}_{1-k}(\mathcal{F}^i_{depth})), \end{aligned} \tag{1}$$

where $\mathcal{F}^i_{rgb}$ and $\mathcal{F}^i_{depth}$ denote the RGB-D features at the ith stage; and $\mathcal{C}at$ denotes the concatenation operation.

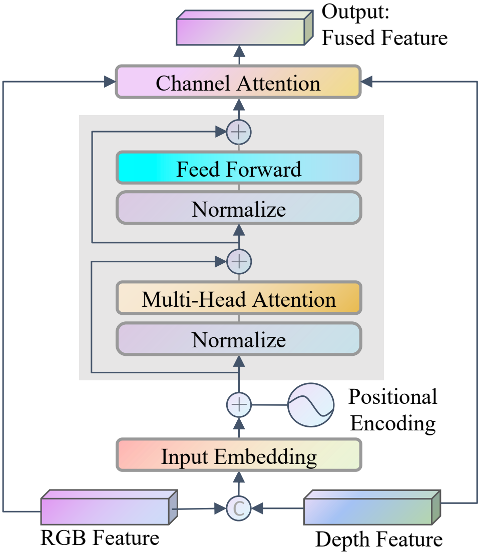

Figure 3: Schematic diagram of the proposed MViT, where C is the concatenation operation, and + is the element-wise addition operation.

3.3 Multi-modal Vision Transformer (MViT)

Recently, Transformer [28] has achieved success on some computer vision tasks (*e.g.*, image classification [9, 27], object detection [6, 32, 35], and video processing [23]). Because of its strong ability to capture global and local attention, we find that Transformer [28] is naturally suitable for the RGB-D feature fusion.

Inspired by ViT [32], we develop its multi-modal variant, MViT, to effectively fuse RGB-D features. Since ViT [32] is proposed for image classification, it introduces a class token to avoid the bias on image representation. Therefore, we remove the class token in the proposed MViT for multi-modal feature fusion. To convert the features extracted from backbones into sequence data, the RGB and Depth features are first flattened in the height h and width w dimensions and then concatenated together. As illustrated in Fig. 3, with feature embedding, the concatenated features $\mathcal{F}_{cat}$ are encoded into d_{hide}-dimensional features $\mathcal{F}_{embed}$. Typically d_{hide} is much smaller than $2hw$ to reduce the computational cost. To distinguish the locations of features, we add positional encoding $\mathcal{P}$ (a set of learnable parameters) to the input embedding. The multi-head attention layer first converts the input features into Query $\mathcal{F}_Q$, Key $\mathcal{F}_K$, and Value $\mathcal{F}_V$, and then the attention output $\mathcal{F}_{attn}$ is computed as

$$\mathcal{F}_{attn} = \sigma(\frac{\mathcal{F}_Q \times \mathcal{F}_K^T}{\sqrt{d_k}}) \times \mathcal{F}_V, \tag{2}$$

where σ represents the Softmax function; d_k denotes the dimension of $\mathcal{F}_K$; × denotes the outer product; and ^{T} indicates the transpose operation.

Different from ViT [32], we add a short-cut connection between the input feature and the Transformer output, and fuse them with a channel attention layer, to improve the feature representation ability. The MViT can be formulated as

$$\mathcal{F}_{fuse}^i = \mathcal{ChnAttn}(\mathcal{Cat}(\mathcal{F}_{attn}^i, \mathcal{F}_{rgb}^i, \mathcal{F}_{depth}^i)), \tag{3}$$

where $\mathcal{F}_{rgb}$ and $\mathcal{F}_{depth}$ denote the input RGB and depth features, respectively; $\mathcal{C}hn\mathcal{A}ttn$ indicates the channel attention layer in Fig. 3; $\mathcal{C}at$ denotes the concatenation operation; and $\mathcal{F}^{i}_{fuse}$ represents the fused features at the ith level of MViT.

Furthermore, we introduce a Bottom-Up pathway between MViTs at different levels, to enrich high-level semantic features with low-level geometric clues, as illustrated in Fig. 1.The Bottom-Up pathway can be formulated as

$$\mathcal{F}^{i+1}_{fuse} = \mathcal{U}p(\mathcal{F}^{i}_{fuse}) + \mathcal{F}^{i+1}_{fuse}, \tag{4}$$

where $\mathcal{U}p$ indicates the upsampling operation.

4 Experimental Results

We evaluate our model on SUN RGB-D [26] and NYU Depth v2 [25], which contain 10,335 and 1,449 RGB-D images, respectively. The training-test splits keep the same as official. Mean average precision (*mAP*) and average precision (*AP*) are adopted as evaluation metrics, which are the same as those proposed by PASCAL VOC [11].

4.1 Implementation Details

We implement our model with the MMDetection toolbox [7] based on PyTorch. Faster R-CNN head proposed by Ren *et al.* [24] is adopted as the classification and regression head. The proposed MViT is set to have depth D as 2 and head H as 2 to learn global and local attention, respectively. For fair comparison, VGG-16 and VGG-11 pre-trained on ImageNet [8] are the default backbones to extract features from RGB-D images, respectively. Following [18, 31], we only work with 19 major furniture categories available in the two datasets: bathtub, bed, bookshelf, box, chair, counter, desk, door, dresser, garbage bin, lamp, monitor, nightstand, pillow, sink, sofa, table, television, and toilet.

On SUN RGB-D [26], models are trained with stochastic gradient descent (SGD) optimizer with initial learning rate as 0.01 and batch size as 4. We adopt the linear warm-up strategy with 500 warm-up iterations. Weight decay and momentum are set as 0.0001 and 0.9, respectively. During training, images are horizontally flipped with a probability of 0.5 for data augmentation. The input images are resized to 608 × 800. All models are trained with 24 epochs. Same as [18, 31], we finetune the SUN RGB-D [26] pre-trained models on NYU Depth v2 [25] with learning rate 0.001.

4.2 Results on SUN RGB-D and NYU Depth v2

We compare the proposed FETNet against recent state-of-the-art RGB-D object detection methods. For these methods, we adopt the results reported in their papers.

As shown in Table 1, FETNet achieves the best performance and promotes *mAP* to 54.5 on SUN RGB-D [26], surpassing state-of-the-art RGB-D detectors at least **7.0%**. The proposed FETNet significantly improves the performance on chair, counter, desk, door, garbage bin, lamp, monitor, nightstand, pillow, sink, and toilet. As a multi-class classification and regression task, the significant improvement in these categories may lead to a slight decline in the remaining categories due to the inter-class balance.

Table 2 shows the detection results on NYU Depth v2 [25] of the pre-trained models. FETNet boosts *mAP* to 54.0, substantially surpassing all the RGB-D detectors. In detail,

Method	***mAP***	*bathhub*	*bed*	*bookshelf*	*box*	*chair*	*counter*	*desk*	*door*	*dresser*	*garbagebin*
		lamp	*monitor*	*nightstand*	*pillow*	*sink*	*sofa*	*table*	*tv*	*toilet*	
RGBD R-CNN [15]	35.2	49.6	76.0	35.0	5.8	41.2	8.1	16.6	4.2	31.4	46.8
		22.0	10.8	37.2	16.5	41.9	42.2	43.0	32.9	69.8	
Super Transfer [16]	43.8	65.3	83.0	54.4	14.4	46.9	14.6	23.9	15.3	41.3	51.0
		32.1	36.8	46.6	23.4	43.9	61.3	48.7	50.5	79.4	
AC-CNN [19]	**45.4**	65.8	83.3	56.2	16.4	47.5	16.0	24.9	16.6	42.7	53.4
		33.8	39.5	47.1	25.2	45.3	61.9	49.0	54.1	84.2	
CMAC [18]	**47.5**	69.0	86.1	57.9	18.2	50.3	17.4	26.8	17.3	44.4	54.4
		35.6	40.5	49.8	26.7	46.6	67.2	52.9	56.7	84.9	
FETNet (Ours)	**54.5**	62.5	80.9	47.9	13.3	69.3	49.2	30.4	52.6	41.9	56.9
		65.0	43.1	62.0	63.9	65.4	56.3	49.5	40.3	85.5	

Table 1: Experimental results on SUN RGB-D. The results in **red** and **blue** represent the first and second best performances.

Method	***mAP***	*bathhub*	*bed*	*bookshelf*	*box*	*chair*	*counter*	*desk*	*door*	*dresser*	*garbagebin*
		lamp	*monitor*	*nightstand*	*pillow*	*sink*	*sofa*	*table*	*tv*	*toilet*	
RGBD R-CNN [15]	32.5	22.9	66.5	21.8	3.0	40.8	37.6	10.2	20.5	26.2	37.6
		29.3	43.4	39.5	37.4	24.2	42.8	24.3	37.2	53.0	
Super Transfer [16]	49.1	50.6	81.0	52.6	5.4	53.0	56.1	21.0	34.6	57.9	46.2
		42.5	62.9	54.7	49.1	50.0	65.9	31.9	50.1	68.0	
AC-CNN [19]	**50.2**	52.2	82.4	52.5	8.6	54.8	57.3	22.7	34.1	58.1	46.5
		42.9	63.6	55.2	49.7	51.4	66.8	33.5	51.8	70.4	
CMAC [18]	**52.3**	55.6	83.9	54.0	9.8	55.4	59.2	24.1	36.3	58.5	47.2
		45.0	65.8	57.6	52.7	53.8	69.1	35.0	56.9	74.7	
FETNet (Ours)	**54.0**	56.4	78.3	57.3	8.0	68.2	37.6	32.5	44.2	59.1	51.9
		50.8	69.5	59.0	60.8	60.3	69.0	36.0	55.4	71.2	

Table 2: Experimental results on NYU Depth v2. The results in **red** and **blue** represent the first and second best performances.

FETNet promotes the performance on bathtub, bookshelf, chair, desk, door, dresser, garbage bin, lamp, monitor, nightstand, pillow, sink, and table.

The outstanding performance on both datasets indicates the effectiveness of our FETNet.

4.3 Compared with RGB-based Detectors

Method	Reference	Input Modality	Backbone	Inference GFLOPs	mAP
GFLv2 [20]	CVPR 2021	RGB	ResNet-152	169.9	49.8
Ours + GFLv2 [20]		RGB-D	ResNet-50	160.6	**53.7**$^{(+3.9)}$
ATSS [34]	CVPR 2020	RGB	ResNet-152	168.4	50.9
Ours + ATSS [34]		RGB-D	ResNet-50	159.1	**54.3**$^{(+3.4)}$
Dynamic R-CNN [33]	ECCV 2020	RGB	ResNet-152	140.5	53.9
Ours + Dynamic R-CNN [33]		RGB-D	ResNet-50	130.7	**57.2**$^{(+3.3)}$
SABL [29]	ECCV 2020	RGB	ResNet-152	347.1	53.1
Ours + SABL [29]		RGB-D	ResNet-50	337.3	**55.9**$^{(+2.8)}$
Cascade R-CNN [4]	CVPR 2018	RGB	ResNet-152	168.3	53.4
Ours + Cascade R-CNN [4]		RGB-D	ResNet-50	158.5	**56.1**$^{(+2.7)}$
Faster R-CNN [24]	NeurIPS 2015	RGB	ResNet-152	140.5	54.8
Ours + Faster R-CNN [24]		RGB-D	ResNet-50	130.7	**57.9**$^{(+3.1)}$

Table 3: Comparison with state-of-the-art RGB-based object detectors on SUN RGB-D.

Since neither FEM nor MViT relies on specific object detection methods, these two modules can be integrated into existing RGB object detection frameworks to get their corresponding RGB-D object detection extensions. Table 3 illustrates the performance of several common detectors and their RGB-D counterparts on SUN RGB-D [26]. For fair comparison (similar computational complicity), we set RGB-based methods with a deeper backbone network (*i.e.*, ResNet-152) and RGB-D methods with two ResNet-50 to extract features from the two input modalities.

As shown in Table 3, the proposed modules (*i.e.*, FEM and MViT) can steadily improve the detection performance on both single-stage detectors (*i.e.*, GFLv2 [20], ATSS [34]) and two-stage detectors (*i.e.*, Dynamic R-CNN [33], SABL [29], Cascade R-CNN [4], Faster R-CNN [24]), with lower computational complicity.

4.4 Ablation Studies

We execute extensive ablation studies to reveal the characteristics of the proposed method. Since NYU Depth v2 [25] is a subset of SUN RGB-D [26], only the latter is used here.

Method	MFP	FEM	MViT	GFLOPs	mAP
baseline				237.1	41.8
baseline + MFP	✓			257.0	49.9
baseline + MFP + FEM	✓	✓		257.0	53.0
baseline + MFP + MViT	✓		✓	279.3	53.8
FETNet	✓	✓	✓	279.3	**54.5**

Table 4: Ablation study on the proposed FEM and MViT on SUN RGB-D. MFP: the Multi-modal Feature Pyramid layer illustrated in Fig. 1.

Table 4 illustrates the performance gain and computational cost of each component of the proposed FETNet. The baseline method is extended from Faster R-CNN [24]. RGB-D features extracted from backbones are fused by directly element-wise addition. Only the final output feature of the backbone is used for object detection. It achieves 41.8 *mAP*, which is lower than existing methods [16, 18, 19]. With the help of the Multi-modal Feature Pyramid layer (*i.e.*, MFP), implemented with two FPN [22] to aggregate multi-scale features, the detector achieves 49.9 *mAP*. When further integrating the proposed FEM, the performance is boosted to 53.0 *mAP*. It is worth noting that the introduction of FEM did not lead to an increase in computational cost, with significant performance gain. When adopting the proposed MViT, the baseline method is improved to 53.8 *mAP*. This demonstrates that both modules are effective for RGB-D object detection. Furthermore, once we combine these two modules to build FETNet, it can surpass the baseline method with a large margin, which is the best result exceeding existing methods by 7.0%.

We study the exchange proportion and stage in the proposed FEM. As shown in Fig. 4, the cyan dotted line indicates the detection capacity without FEM (*i.e.*, baseline+MFP+MViT), and the magenta line denotes the performance under different exchange proportions. As can be seen, the performance reaches the peak when 1/8 of channels are exchanged. When swapping at 1/2 and larger ratios, the performance is worse than excluding the swap operation. When exchanging at 1/32 channels, it achieves slightly better performance than excluding feature exchanging. Since swapping a large proportion of channels may harm the spatial feature learning ability of the backbone network, while swapping a small proportion is insufficient for information exchange. Exchanging features at different backbone stages is also an important factor. Exchanging features at earlier stages leads to better performance, as it enables the information flow at an earlier stage. Moreover, swapping RGB-D features at more stages has a positive effect on the detector, since features extracted at different stages provide diverse receptive fields. Extensive experiments show that exchanging the RGB-D features at all backbone stages achieves the best performance.

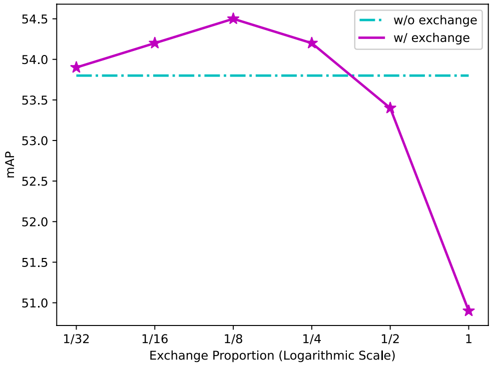

Figure 4: Performance on exchanging different proportion of channels in the proposed Feature Exchange Module on SUN RGB-D.

Method	SCA	BUP	mAP
baseline + MFP + FEM + ViT			53.2
FETNet w/o BUP	✓		53.9
FETNet w/o SCA		✓	53.8
FETNet	✓	✓	**54.5**

Table 5: Ablation study on the proposed Multi-modal Vision Transformer on SUN RGB-D. SCA: the channel attention on the short-cut connection. BUP: the Bottom-Up pathway on different level MViTs.

We compare the proposed MViT with two convolution-based attention methods, to validate the effectiveness of long-range attention for RGB-D feature fusion. We replace the MViT with CBAM [30] and GCBlock [5], the performance drops from 54.5 to 52.3 and 53.1, respectively. The main reason is the corresponding pixels of the two modalities represent different information (*i.e.*, color and distance). Global receptive-field helps the network to understand each modality better. Another reason is the RGB and Depth pairs are not perfectly aligned. This misalignment leads to the inefficiency of convolution-based local attention methods.

We further reveal the characteristics of the proposed MViT. As illustrated in Table 5, the short-cut channel attention connection and the bottom-up pathway both improve the detection capability and cooperate to get the best performance. As for the depth D and head H of the Vision Transformer, the detection capability increases with D and H. Even when we set the depth and head as 1, the performance (54.3 *mAP*) still surpasses existing methods, indicating that the proposed MViT can effectively fuse the features of two modalities.

Method	depth format	mAP
FETNet	HHA [15]	54.0
FETNet	raw depth	**54.5**

Table 6: Performance of different depth formats (*i.e.*, HHA and raw depth) of the FETNet on SUN RGB-D.

We finally execute experiments on two widely-used formats of depth information. As illustrated in Table 6, the performance of the two formats is similar. The proposed FETNet is robust for the format of depth data. It proves that FETNet can effectively extract and fuse multi-modal features from raw depth data, and thus handcrafted conversion (*i.e.*, HHA [15]) becomes not necessary.

5 Conclusion

In this paper, we propose a Feature Exchange Transformer Network (FETNet) to fuse multimodal features and detect objects effectively. Two well-designed components are introduced. The FEM module is introduced to exchange part of the features extracted at each backbone stage. It adds an information flow and bridges the gap between the RGB-D features. The MViT module is developed to effectively exploit the global and local attention between the two modalities. Extensive experiments show our FETNet surpasses state-of-the-art detectors with a large margin in the RGB-D object detection.

References

[1] Manuel Blum, Jost Tobias Springenberg, Jan Wülfing, and Martin Riedmiller. A learned feature descriptor for object recognition in RGB-D data. In *IEEE International Conference on Robotics and Automation*, pages 1298–1303. IEEE, 2012.

[2] Liefeng Bo, Kevin Lai, Xiaofeng Ren, and Dieter Fox. Object recognition with hierarchical kernel descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1729–1736, 2011.

[3] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Depth kernel descriptors for object recognition. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 821–826. IEEE, 2011.

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade R-CNN: Delving into high quality object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[5] Yue Cao, Jiarui Xu, Stephen Lin, Fangyun Wei, and Han Hu. Gcnet: Non-local networks meet squeeze-excitation networks and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015.

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[14] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

[15] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360. Springer, 2014.

[16] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2827–2836, 2016.

[17] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In *Asian Conference on Computer Vision*, pages 213–228. Springer, 2016.

[18] Guanbin Li, Yukang Gan, Hejun Wu, Nong Xiao, and Liang Lin. Cross-modal attentional context learning for RGB-D object detection. *IEEE Transactions on Image Processing*, 28(4):1591–1601, 2018.

[19] Jianan Li, Yunchao Wei, Xiaodan Liang, Jian Dong, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Attentive contexts for object detection. *IEEE Transactions on Multimedia*, 19(5):944–954, 2016.

[20] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. *arXiv preprint arXiv:2011.12885*, 2020.

[21] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[23] Zhouyong Liu, Shun Luo, Wubin Li, Jingben Lu, Yufan Wu, Chunguo Li, and Luxi Yang. Convtransformer: A convolutional transformer network for video frame synthesis. *arXiv preprint arXiv:2011.10185*, 2020.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[25] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[26] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015.

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[29] Jiaqi Wang, Wenwei Zhang, Yuhang Cao, Kai Chen, Jiangmiao Pang, Tao Gong, Jianping Shi, Chen Change Loy, and Dahua Lin. Side-aware boundary localization for more precise object detection. In *European Conference on Computer Vision*, pages 403–419. Springer, Cham, 2020.

[30] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[31] Xiangyang Xu, Yuncheng Li, Gangshan Wu, and Jiebo Luo. Multi-modal deep feature learning for RGB-D object detection. *Pattern Recognition*, 72:300–313, 2017.

[32] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *European Conference on Computer Vision*, pages 323–339. Springer, 2020.

[33] Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic R-CNN: Towards high quality object detection via dynamic training. *arXiv preprint arXiv:2004.06002*, 2020.

[34] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9759–9768, 2020.

[35] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.